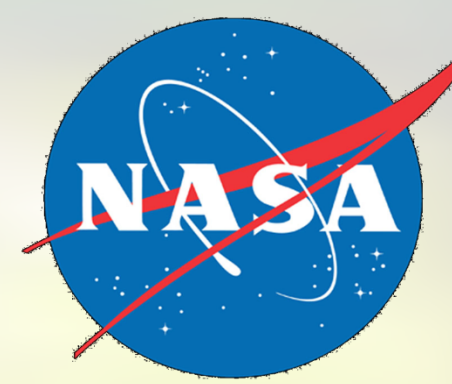




Science Data Preservation: Implementation and Why it is Important

Steven Kempler¹, John Moses¹, Irina Gerasimov², James Johnson², Bruce Vollmer¹ Michael Theobald², Dana Ostrenga², Suraiya Ahmad², Hampapuram Ramapriyan¹, Mohammad Khayat²

Steven.J.Kempler@nasa.gov

¹NASA Goddard Space Flight Center, ²NASA Goddard Space Flight Center/SES DA3

Abstract

Remote Sensing data generation by NASA to study Earth's geophysical processes was initiated in 1960 with the launch of the first Television Infrared Observation Satellite Program (TIROS), to develop a meteorological satellite information system. What would be deemed as a primitive data set by today's standards, early Earth science missions were the foundation upon which today's remote sensing instruments have built their scientific success, and tomorrow's instruments will yield science not yet imagined. NASA Scientific Data Stewardship requirements have been documented to ensure the long term preservation and usability of remote sensing science data. In recent years, the Federation of Earth Science Information Partners and NASA's Earth Science Data System Working Groups have organized committees that specifically examine standards, processes, and ontologies that can best be employed for the preservation of remote sensing data, supporting documentation, and data provenance information.

This presentation describes the activities, issues, and implementations, guided by the NASA Earth Science Data Preservation Content Specification (423-SPEC-001), for preserving instrument characteristics, and data processing and science information generated for 20 Earth science instruments, spanning 40 years of geophysical measurements, at the NASA's Goddard Earth Sciences Data and Information Services Center (GES DISC). In addition, unanticipated preservation/implementation questions and issues in the implementation process are presented.

Motivation

NASA remote sensing data are a national resource with great scientific value that will be preserved and shared for future scientific research, by generations to come. Maintaining and ensuring their use is essential as we learn new ways to utilize these data in science research and applications. Conscientious preservation preparations will lead to optimal information preservation.

No preservation = Loss of Future Long Term Climate Records

NASA Earth Science Data Preservation Content Specification (423-SPEC-001)

1. Category	2. Content Item	3. Definition/Description
Preflight/Pre-Operations Calibration	Instrument Description	Documentation of Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, instrument geometric calibration (geo-location offsets), noise characteristics, etc.).
	Preflight/Pre-operational Calibration Data	Numeric (digital data) files of Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, instrument geometric calibration (geo-location offsets), noise characteristics, etc.).
Science Data Products	Raw Data and Derived Products	Raw data are data values at full resolution as directly measured by a spaceborne, airborne or <i>in situ</i> instrument. Derived products are higher level products (level 1b through 4) where calibration and geo-location transformations have been applied to generate sensor units, and/or algorithms have been applied to generate gridded geophysical parameters.
	Metadata	Information about data to facilitate discovery, search, access, understanding and usage associated with each of the data products.
Science Data Product Documentation	Product Team	Names of key science team leads and product team members (development, help desk and operations), roles, performing organization, contact information, sponsoring agencies or organizations and comments about the products.
	Product Requirements	Requirements and designs for each science data product, either explicitly or by reference to the requirements/design documents. Product requirements and designs should include content, format, latency, accuracy and quality.
	Processing and Algorithm Version History	For all products held in the archive, documentation of processing history and production version history, indicating which versions were used when, why different versions came about, and what the improvements were from version to version. For all products held in the archive, the versions of source code used to produce the products should be available at the archive.
	Product Generation Algorithm	Detailed discussion of processing algorithms, outputs, error budgets and limitations. Processing algorithms and their theoretical (scientific and mathematical) basis, including complete description of any sampling or mapping algorithm used in creation of the product, geo-location, radiometric calibration, geophysical parameters, sampling or mapping algorithms used in creation of the product, algorithm software documentation, & high-level data flow diagrams. Description of how the algorithm is numerically implemented.
	Product Quality	Description of the impact to product quality due to issues with computationally intensive operations (e.g., large matrix inversions, truncation and rounding). Documentation of product quality assessment (methods used, assessment summaries for each version of the datasets). Description of embedded data at the granule level including quality flags, product data uncertainty fields, data issues logs, etc. Relevant test reports, reviews, and appraisals.
Mission Data Calibration	Product Application	Useful references to published articles about the use of the data and user feedback received by the science and instrument teams about the products. Includes reports of any peculiarities or notable features observed in the products.
	Calibration Method	The methods used for instrument/sensor radiometric and geometric calibration while in operation (e.g., in orbit). The source code used in applying the calibration algorithms. Documentation of in-line changes to calibration or to instrument or platform operations or conditions that occur throughout the mission.
Science Data Product Software	Calibration Data	Instrument and platform engineering data collected during operations (e.g., in orbit), including platform and instrument environment, events and maneuvers; attitude and ephemeris; aircraft position; acquisition logs that record data gaps; calibration look-up tables; and any significant external event data that may have impacted the observations.
	Science data product generation software and software documentation	Source code used to generate products at all levels in the science data processing system. Software release notes, including references to versions of operating systems, compilers, commercial software libraries used in the code. Versions of science data product software should be archived for each major product release. A major product release is characterized by the appearance of peer reviewed publications where reported results are based on the product version.
Science Data Product Algorithm Inputs	Ancillary data and documentation	Complete information on any ancillary data or other data sets used in generation or calibration of the data set or derived product, either explicitly in data descriptions or by reference to appropriate publications. Ancillary data should be stored with the products unless it is available from another permanent archive facility.
Science Data Product Validation	Datasets and documentation	Accuracy of products, as measured by validation testing, and compared to accuracy requirements. Description of validation process, including identification of validation data sets, measurement protocols, data collection, analysis and accuracy reporting.
Science Data Software Tools	Software and documentation	Product access (reader) tools. Software source code that would facilitate use of the calibration data, ancillary data and the data products at all levels. Includes software source code useful for creating programs that will read and display the calibration data, ancillary data and product data and metadata values. Commercial tools should be identified with appropriate references.

What is Science Data Preservation

What is Stewardship: The responsible overseeing and protection of something considered worth caring for and preserving

What is Preservation: The act of keeping in perfect or unaltered condition; maintain unchanged

The GES DISC, as are all NASA Distributed Active Archive Centers (DAACs), is responsible for stewarding NASA Earth science data and thoroughly preparing the data, and associated documentation, for long term preservation.

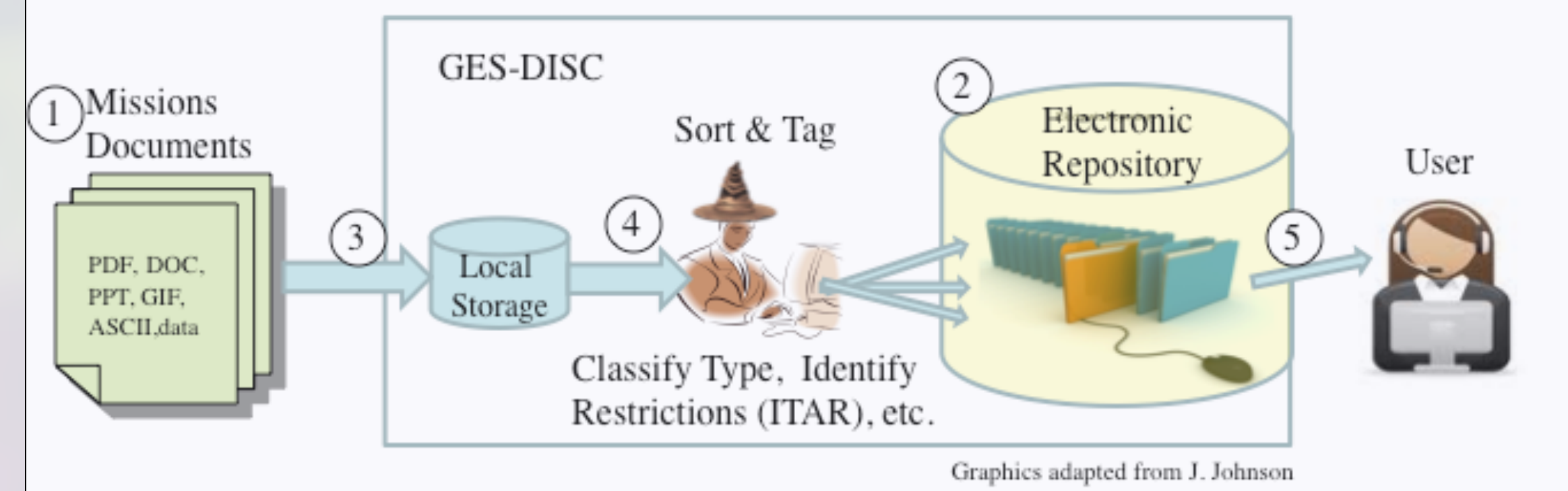
The GES DISC Data Preservation activity goals are to:

- Acquire and safely archive all data that cannot be regenerated if lost, and all mission related documentation that contributes to the description and usage of instrument generated, and science research derived data
- Implement an environment that will ensure the preservation of this information

Implementing Data Preservation

Work Breakdown

- Identify documentation
- Specify and implement preservation environment
- Retrieve documentation
- Archive and catalog documentation
- Implement retrieval and distribution services



Dataset Information to be Preserved

← Details:

- Preflight/Pre-operations Calibration**
- Science Data Products**
- Science Data Products Documentation**
- Mission Data \Calibration**
- Science Data Product Software**
 - Software is for documentation purpose only
- Science Data Product Algorithm Inputs**
- Science Data Product Validation**
- Science Data Software Tools**

Mission Datasets Resident at the GES DISC

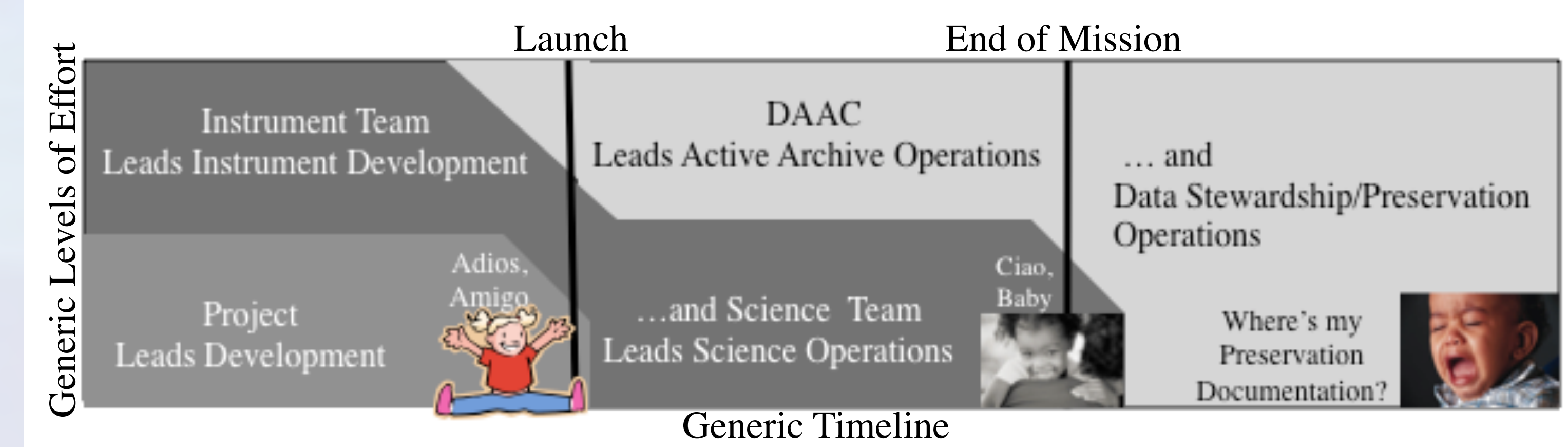
Atmospheric Composition

- Total Ozone Mapping Spectrometer (TOMS)
- Upper Atmosphere Research Satellite (UARS)
- Aura: - Ozone Monitoring Instrument (OMI)
 - Microwave Limb Sounder (MLS)
 - High Resolution Dynamics Limb Sounder (HIRDLs)
- Atmospheric CO2 Observations from Space (ACOS)
- Historical datasets from Nimbus, SME, others
- Coming: Orbiting Carbon Observatory 2 (OCO-2)

Why is Science Data Preservation Important

No preservation = Loss of Future Long Term Climate Records

Early data preservation preparations are essential...



... otherwise mission/science experts move on before their expertise is captured

Data Point: Currently analyzing an inventory of 12,000 reel to reel tapes containing 1960's vintage satellite data, determining data content and quality, with very little documentation.

GES DISC Preservation Implementation Status

- Identify documentation
 - GES DISC Science Support identified specific information needed per mission in Data Preservation Mission List*
- Specify and implement preservation environment
 - Local archive...afterall, documentation is data*
 - Fedora Commons: Population has begun for test datasets (HIRDLs)*
 - Still exploring NASA Technical Reports Server (NTRS) and NASA Aeronautics and Space Database (NA&SD)*
- Retrieve documentation
 - In discussion with each mission science team*
- Archive and catalog documentation
 - Pathfinding with HIRDLs documentation. Other prototyping*
- Implement retrieval and distribution services
 - Concept in steps: 1. Internal Access; 2. External Access (basic retrieval); 3. External Access (add services); 4. Iterate with other DAACs/community*
 - Treat ITAR differently from non-ITAR documentation*

What We Have Learned and What We Have to Resolve (so far)

- Heritage missions have a lot to catch up
- Sensitive vs. non-sensitive information: Looking to extract the ITAR pieces from the bulk of the documentation
- Still examining possible issues with NTRS
- Level of service:
 - For relatively occasional use, what is the right level of services returned per cost?
 - What services should be provided to users seeking seldom used documentation?
 - Is there a reason why users can not access the documentation archive themselves?
- Do we preserve field campaign data? ACOS data? Model data? MEaSUREs data? Validation datasets?
- How to acquire information from science team?
- How is preservation documentation validated?

Precipitation

- Tropical Rainfall Measuring Mission (TRMM)
- Hydrology Data Collections
- Coming: Global Precipitation Mission (GPM)

Modeling

- Global Modeling Assimilation Office (GMAO)
- Global Land Data Assimilation System (GLDAS)
- North American Land Data Assimilation System (NLDAS)

Atmospheric Dynamics

- Aqua: Atmospheric Infrared Sounder (AIRS)
- TIROS Operational Vertical Sounder (TOVS)

Solar Irradiance

- Solar Radiation and Climate Experiment (SORCE)

MEaSUREs Data Sets from 7 Projects